GENERALIZABLE GEOMETRIC IMAGE CAPTION SYNTHESIS

Yue Xin[1]*, **Wenyuan Wang**[2]*, **Rui Pan**[3]*,
Howard Meng[3], **Renjie Pi**[3], **Ruida Wang**[3], **Tong Zhang**[3]
[1]Shanghai Jiao Tong University, [2]Wuhan University, [3]UIUC

ABSTRACT

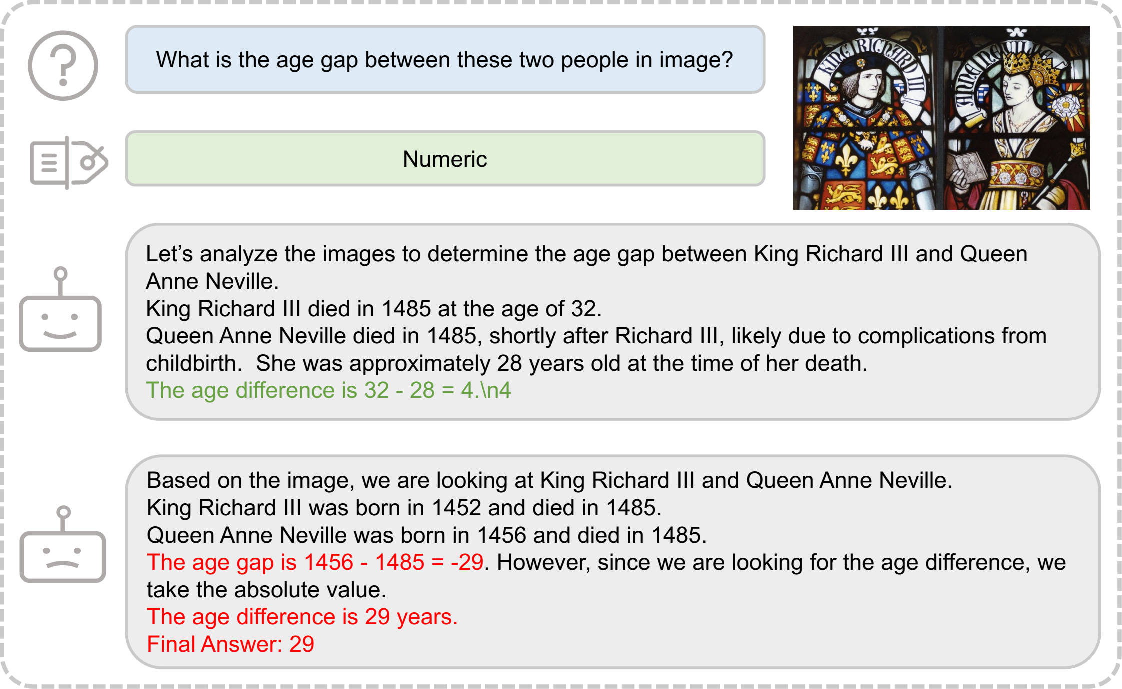
Multimodal large language models have various practical applications that demand strong reasoning abilities. Despite recent advancements, these models still struggle to solve complex geometric problems. A key challenge stems from the lack of high-quality image-text pair datasets for understanding geometric images. Furthermore, most template-based data synthesis pipelines typically fail to generalize to questions beyond their predefined templates. In this paper, we mitigate this issue by introducing a complementary RLHF process into the data generation pipeline. By adopting RAFT to refine captions for image-text pairs generated from approximately 50 templates and using reward signals derived from mathematical problem-solving tasks, our pipeline successfully captures the key features of geometry problem-solving. This enables better task generalization and yields non-trivial improvements. Furthermore, the generated dataset also enhances the general mathematical reasoning capabilities of multimodal large language models beyond the domain of geometric problems, yielding accuracy improvements of 3.1%–5.5% in arithmetic, algebraic, and numerical tasks even with non-geometric input images.

1 Introduction

Multimodal Large Language Models (MLLMs) have exhibited impressive capabilities across a variety of vision-related tasks, including Visual Question Answering (VQA), visual grounding, and image captioning. Recent MLLMs, such as Qwen2.5-VL, Intern2.5-VL, and LLaVA-Next [1, 2, 3], have shown superior performance compared to specialized vision models across a wide range of visual tasks, highlighting the potential of unified multimodal architectures. As the field advances, there has been increasing interest in enhancing the reasoning capabilities of MLLMs [4, 5], which is regarded as a crucial factor in extending the performance boundaries of these models. Among various reasoning tasks, mathematical reasoning [6] has attributed particular attention due to its structured problem-solving nature, offering a clear pathway for MLLMs to develop and improve their reasoning skills.

Research from MathVerse [6] indicates that MLLMs perform best when the input is purely textual, while their performance declines significantly with visual-only inputs. This highlights the urgent need for MLLMs to develop strong cross-modal reasoning capabilities, which involves accurately and comprehensively transferring information from the image to the text. Although numerous geometry and math datasets have been introduced [7, 8, 9] to boost various aspects of model performance, high-quality datasets explicitly designed for cross-modal reasoning remain scarce. That is primarily because in existing datasets, the alignment between images and captions is often asymmetrical. For instance, in geometric problems, two lines of equal length can be easily described textually but may not be correspondingly annotated or visually distinct in the image. Such discrepancies hinder the model's ability to learn robust cross-modal reasoning.

Meanwhile, Reinforcement Learning (RL) has been demonstrated to significantly improve model reasoning and generalization capabilities [10]. Its reward-driven framework is particularly effective for cross-modal reasoning, allowing models to optimize decision-making through interactive feedback [11, 12, 13]. Building on these insights, we employ the RAFT method [14] and design a reward function that incorporates both reasoning and caption rewards. This facilitates the alternating optimization of dataset quality and model reasoning abilities, leading to improved results on complex multimodal tasks.



What is the age gap between these two people in image?

Numeric

Let's analyze the images to determine the age gap between King Richard III and Queen Anne Neville.
King Richard III died in 1485 at the age of 32.
Queen Anne Neville died in 1485, shortly after Richard III, likely due to complications from childbirth. She was approximately 28 years old at the time of her death.
The age difference is 32 - 28 = 4.\n4

Based on the image, we are looking at King Richard III and Queen Anne Neville.
King Richard III was born in 1452 and died in 1485.
Queen Anne Neville was born in 1456 and died in 1485.
The age gap is 1456 - 1485 = -29. However, since we are looking for the age difference, we take the absolute value.
The age difference is 29 years.
Final Answer: 29

Figure 1: MLLMs learn from our synthetic geometric mathematical problems and generalize to algebraic cases with even non-geometric input images.

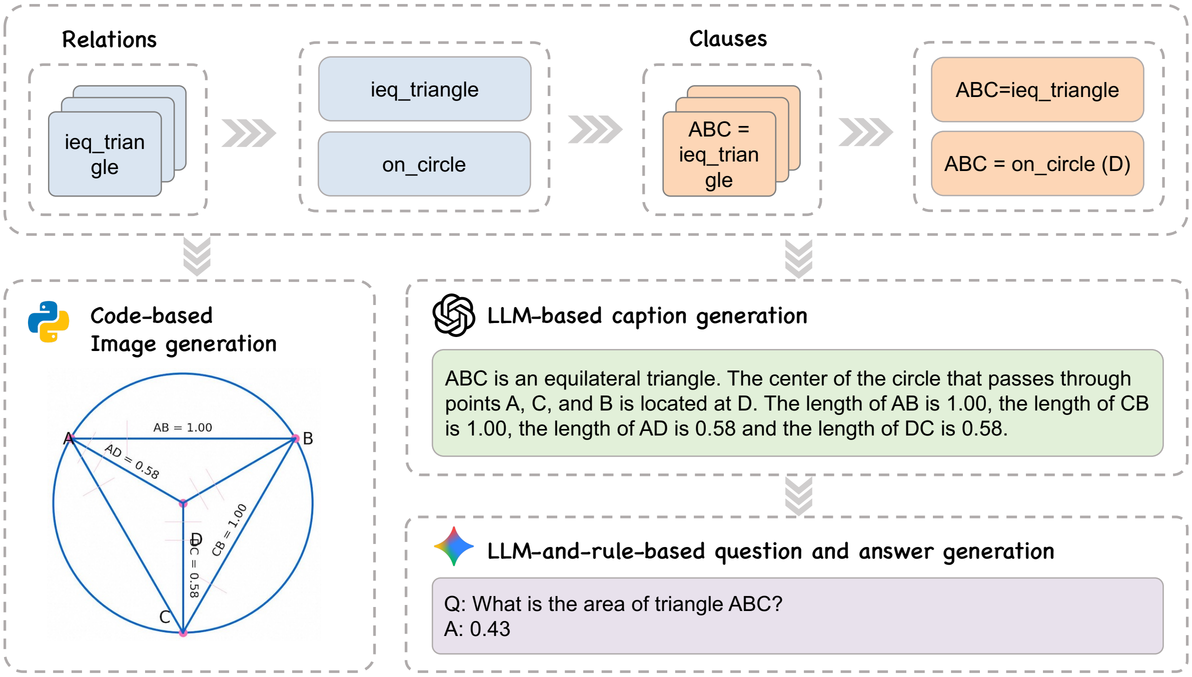
To bridge the gap between visual and linguistic modalities, we propose an RL-based data refinement engine that iteratively enhances data quality. Utilizing this pipeline, we introduce a novel geometry dataset comprising 10,000 image-caption pairs. To the best of our knowledge, this is the first high-quality dataset in which visual and textual information are fully aligned, making it a valuable resource for improving cross-modal reasoning. Experimental results demonstrate that our dataset significantly enhances models' cross-modal reasoning abilities and their performance on geometric image textualization tasks. Furthermore, models trained on our dataset exhibit strong generalization capabilities on other mathematics-focused benchmarks, including MathVerse and MathVista. In summary, our main contributions are summarized as follows:

- We introduce **GeoReasoning-10K**, a new dataset containing 10,000 carefully constructed image-caption pairs where visual and textual information are fully equivalent. This dataset serves as a high-quality resource for training cross-modal reasoning models.
- We propose **Geo-Image-Textualization**, a scalable RL-based framework for generating and refining high-quality synthetic image-caption pairs in geometry. Our iterative RL-driven optimization significantly enhances data alignment and semantic accuracy, and demonstrates generalization to out-of-domain geometric tasks.
- Extensive experiments show that the improvements facilitated by GeoReasoning extend beyond geometric tasks to non-geometric mathematical tasks and even to non-mathematical domains such as art and engineering.

2 Related Works

2.1 Data Generation

Recent studies have highlighted the scarcity of high-quality geometry image–caption datasets, which limits fine-grained cross-modal reasoning in geometric tasks. Following AlphaGeometry [15], Autogeo [16] proposed an automatic generation engine to produce image-caption pairs, constructing a 100K dataset named AutoGeo-100k. MATHGLANCE [17] introduced GeoPeP, a perception-oriented dataset of 200K structured geometry image-text pairs explicitly annotated with geometric primitives and spatial relationships. MagicGeo [18] formulates diagram synthesis as a coordinate optimization problem, ensuring formal geometric correctness via solvers before coordinate-aware text generation.



ABC is an equilateral triangle. The center of the circle that passes through points A, C, and B is located at D. The length of AB is 1.00, the length of CB is 1.00, the length of AD is 0.58 and the length of DC is 0.58.

Q: What is the area of triangle ABC?
A: 0.43

Figure 2: The geometry data synthesis pipeline, where a graph-based representation similar to AutoGeo [16] is employed for generating the final geometry images. The relation library comprises over 50 basic geometric relationships that can be composed into complex ones, providing comprehensive coverage for geometric problems of various difficulties. The image-caption pair is utilized for the SFT stage, while the caption-QA pair for the RLVR stage.

Despite the advances, existing pipelines still struggle to guarantee full modality alignment, i.e., captions frequently omit visual details, while images lack exhaustively aligned textual descriptions.

2.2 Image Captioning

Image captioning aims to generate comprehensive descriptions of visual content and has been widely studied for natural images. While general-purpose multimodal large language models (MLLMs) such as mPLUG-Owl2 [19], MiniGPT-4 [20], and BLIP-3 [21] can perform captioning to some extent, their effectiveness is often limited by insufficient fine-grained cross-modal alignment. Image-Textualization [22] mitigates this issue by integrating multiple vision experts to produce more detailed and accurate captions.

However, the potential of utilizing image captioning to enhance geometric reasoning capacity remains underexplored. OmniCaptioner [23] proposes a unified visual captioning framework that converts diverse images into fine-grained textual descriptions. Nonetheless, its geometric annotations are derived from AutoGeo and MAVIS, largely relying on synthetic or loosely aligned pairs rather than fully equivalent visual-textual representations. Moreover, the scarcity of high-quality geometric image-caption pairs makes it difficult to accurately extract and align geometric information. As a result, current models underperform on geometric image textualization compared to natural image captioning and general visual reasoning benchmarks.

3 Methods

In this section, we introduce our Geo-Image-Textualization data generation pipeline first, followed by our RAFT method used for data refinement.

3.1 Geo-Image-Textualization Data Generation Engine

The proposed data generation pipeline mainly contains three parts: the relation sampling, image-caption pair generation, and question-answer generation procedure, as shown in Figure. 2.

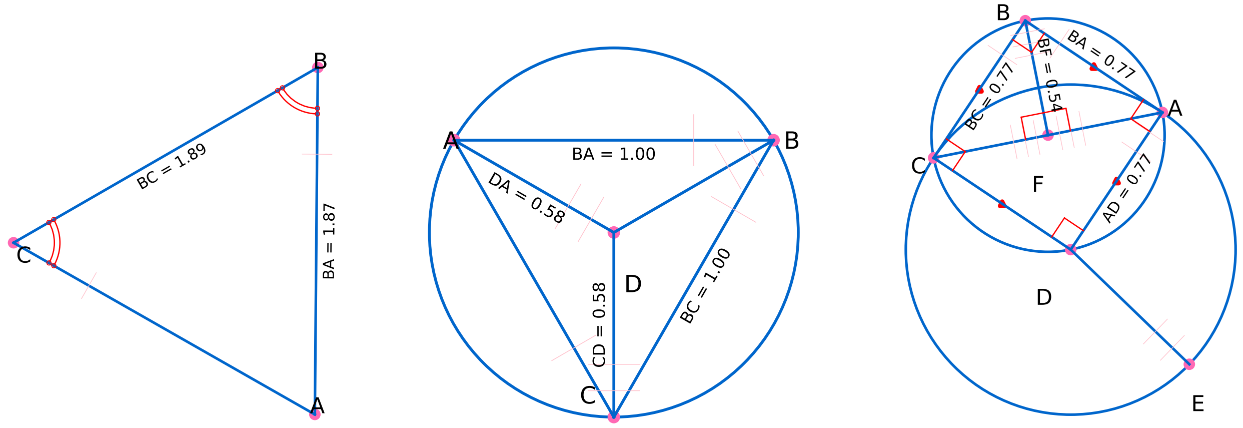


Figure 3: These geometry problems are composed from templates in our relation library, corresponding to easy, medium, and hard difficulty levels, respectively, where the pink ticks and red arcs indicate equal-length segments and equal angles. For visual clarity, this figure has modified colors, font sizes, and line thicknesses compared to the original images in our constructed dataset; please refer to the original dataset for precise details.

3.1.1 Relation Sampling

We develop the Geo-Image-Textualization pipeline upon the data generation procedure in AlphaGeometry. In our framework, *Relations* act as fundamental construction operations that systematically generate diverse yet semantically coherent geometric premises for subsequent theorem synthesis. Each relation (e.g., `angle_mirror`, `circumcenter`, etc.) encodes a precise geometric procedure—such as reflecting a point across an angle bisector or locating the circumcenter of a triangle. In addition to the construction rule, each definition maintains dependency metadata, specifying which primitive objects (points, lines, circles) and prior constructions it depends on. This enables the symbolic engine to get the minimal set of premises required to derive a given theorem.

After sampling relations, each relation is converted into a clause, with associated point variables. For instance, the relation `angle_mirror x a b c` denotes: given points `a`, `b`, and `c`, construct point `x` as the reflection of `c` across angle $\angle ABC$. Finally, the system constructs a graph-based representation in AutoGeo [16] to model geometric problems. Each clause, corresponding to either a geometric construction or a relational assertion, is incorporated into the graph by instantiating nodes for geometric entities (e.g., points, lines, circles) and establishing edges that encode their interdependencies. Before adding each clause, the system verifies the logical correctness of the selected set of predefined geometric rules.

3.1.2 Image-Caption Pair Generation

The geometric graph encodes all relevant entities, including points, lines, and circles, enabling the straightforward rendering of basic geometric elements, similar to AutoGeo. However, a fundamental limitation of AutoGeo is that the captions cannot be directly inferred from the rendered images because the visual content and the textual description are not semantically aligned. We argue that this misalignment constitutes a critical bottleneck in cross-modal reasoning.

To address this issue, we introduce a set of visual augmentation strategies that explicitly encode semantic relationships within the image, mainly including the following properties, as shown in Figure. 2.

1. **Segment Equality Representation:** We use short perpendicular ticks to indicate equal-length line segments. When multiple pairs of equal segments exist, we distinguish them using a different number of ticks (e.g., one tick, two ticks).
2. **Angle Annotations:** For angles that are integer multiples of 15° within the range $[15°, 165°]$, we directly annotate the angle values within the image.
3. **Parallel and Perpendicular Indicators:** Parallel lines are marked using matching directional triangles, and right angles are indicated using a small square symbol at the vertex.
4. **Equal Angle Representation:** Equal angles are denoted by marking them with the same number of arcs, consistent with conventional geometric notation.
5. **Intersection and Collinearity:** Dashed lines are used to explicitly indicate intersections and collinearity relationships among points or segments.

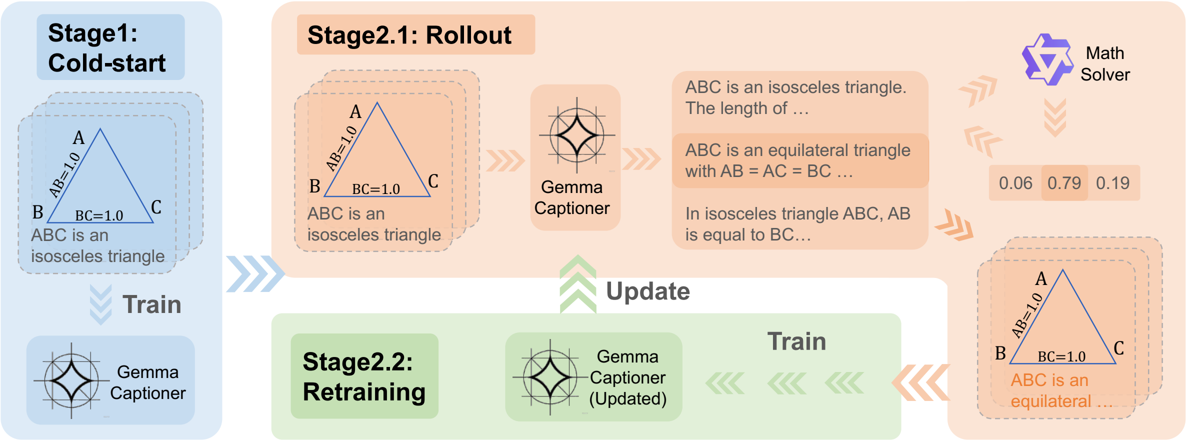

Stage1: Cold-start
A
B
C
AB=1.0
BC=1.0
ABC is an isosceles triangle
Train
Gemma Captioner
Stage2.1: Rollout
ABC is an isosceles triangle
Gemma Captioner
ABC is an isosceles triangle. The length of …
ABC is an equilateral triangle with AB = AC = BC …
In isosceles triangle ABC, AB is equal to BC…
Math Solver
0.06
0.79
0.19
Update
ABC is an equilateral …
Stage2.2: Retraining
Gemma Captioner (Updated)
Train

Figure 4: The workflow of the RAFT method. In Stage 1, the model is trained to develop a preliminary ability to generate image captions. In Stage 2, an alternating optimization strategy is employed to jointly refine the generated captions and enhance the model's overall performance. The data of Stage 1 comes from the rule-based image-caption generation pipeline illustrated in Figure 2.

For each clause in the symbolic representation, we apply a refined, rule-based template to convert it into natural language. These captions accurately describe the geometric diagram, including object relationships, special angle values, and other visual properties. Additionally, the captions explicitly state the lengths of specific line segments when such information is visually annotated in the image. By ensuring that all semantic content present in the image is mirrored in the caption, we achieve full cross-modal alignment.

Building upon these strategies, we construct a comprehensive image-caption generation engine. The engine is entirely rule-based, offering a fast, rigorous, reliable, and cost-effective solution for producing high-quality image-caption pairs to serve as a solid foundation for downstream multi-modal reasoning tasks.

3.2 Question-Answer Pair Generation

The most fundamental requirements for generating questions lie in three aspects. First, the generated question should be based on the caption, i.e., should not be irrelevant to the caption. Second, any information already present in the caption should be removed, as this would dilute the impact of the caption on the model's correctness in subsequent RLVR steps and lead to reward ambiguity. Last, the question should be compatible with the given information, so that it can be logically answered based on what is provided.

Based on these requirements, we propose a rule-and-LLM-based pipeline to systematically generate the question and answer based on the pre-generated caption. Specifically, we first design a prompt that satisfies all the above conditions, using a relatively low temperature (0.2 in our experiments) to encourage the large model (Gemini 2.5 Flash) to generate initial questions based on the caption, while also labeling those that are inconsistent with the caption. For the inconsistent questions, we then switch to a different prompt, encouraging the model to incorporate additional information and formulate new questions accordingly, while increasing the temperature to 0.8. This process continues until a self-consistent question is generated for the first time. The detailed two-stage prompt design is provided in Appendix A.

3.3 RAFT Framework for Data Refinement

Our proposed RAFT framework iteratively optimizes both the model and the dataset through a novel alternating paradigm. The method consists of two phases: (1) a cold-start phase to bootstrap initial captioning capabilities, and (2) a RAFT [14] optimization phase that cyclically refines the dataset and model via reinforcement learning. The overall framework is shown in Figure 5.

3.3.1 Cold-Start Phase

To initialize the model's ability to generate geometrically aligned captions, we first perform supervised fine-tuning (SFT) on the base vision language model using the GeoReasoning-10K dataset. This phase minimizes the standard

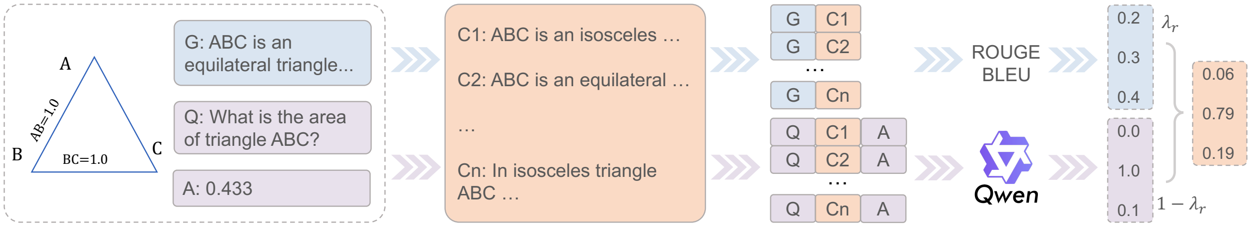

A
B
C
AB=1.0
BC=1.0
G: ABC is an equilateral triangle...
Q: What is the area of triangle ABC?
A: 0.433
C1: ABC is an isosceles …
C2: ABC is an equilateral …
…
Cn: In isosceles triangle ABC …
G C1
G C2
…
G Cn
Q C1 A
Q C2 A
…
Q Cn A
ROUGE BLEU
Qwen
0.2
0.3
0.4
λ_r
0.0
1.0
0.1
$1-\lambda_r$
0.06
0.79
0.19

Figure 5: The left square demonstrates a sample, where 'G', 'Q', 'A' indicate the originally rule-based generated caption, the corresponding question, and the answer, respectively. The rewarding process has two reward functions, the caption reward and the reasoning reward. The former is a weighted sum of the ROUGE and BLEU scores. For the latter, we construct a series of prompts by concatenating each pre-generated question with its corresponding candidate caption. These prompts are then fed into a language model to generate answers, which are compared against the ground truth to determine correctness to get the correctness reward. Then we extract the format tags to get the format reward. The reasoning reward is then defined as the weighted sum of the correctness reward and the format reward, which can be formally expressed as: $R(c, I) = \lambda_r \cdot R_{\text{reasoning}}(c, q) + (1 - \lambda_r) \cdot R_{\text{caption}}(c, c^\star)$, where λ_r indicates the weight of the reasoning reward.

cross-entropy loss:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(I,c^\star)\sim\mathcal{D},}[\log P_{\theta_0}(c|I)] \tag{1}$$

where I denotes an input geometric image, $c^\star$ and c indicate the ground-truth caption and the predicted caption respectively, and D_0 represents the initial dataset. The model parameter θ_0 is optimized to establish basic image-to-text mapping capabilities.

3.3.2 RAFT Optimization Phase

The RAFT optimization phase operates in alternating stages, as shown in Figure 5.

Rollout Experience Generation Suppose the current iteration is t. For each image I in the current dataset $\mathcal{D}_t$, we first generate N candidate captions $\{c_i\}(i = 1, 2, \cdots, N)$ using the current vision language model with parameter θ_t. Then, we utilize a specifically designed reward function $R(c_i, Q_i, c^\star)$ (detailed introduced in Section 3.3.3) to score each caption. Last, we retain the top-K caption $c_{\text{best}} = \arg\max_{c_i} R(c_i, Q_i, c^\star)$ to update the current dataset and construct the refined dataset $\mathcal{D}_{t+1}$.

Model retraining We update the model by training on $\mathcal{D}_{t+1}$ for one epoch, which is:

$$\theta_{t+1} = \arg\max_{\theta_t} \mathbb{E}_{(I,c_{\text{best}})\sim\mathcal{D}_{t+1}}[\log P_{\theta_t}(c_{\text{best}}|I)] \tag{2}$$

This iterative process continues for $T = 5$ epochs, progressively enhancing both dataset quality and model performance.

3.3.3 Reward function

The composite reward $R(c, q, c^\star)$ balances task correctness and caption-image alignment, as shown in Eq. 3:

$$R(c, I) = \lambda_r \cdot R_{\text{reasoning}}(c, q) + (1 - \lambda_r) \cdot R_{\text{caption}}(c, c^\star) \tag{3}$$

Reasoning reaward To evaluate a candidate caption's utility for solving downstream tasks, we leverage a frozen large language model of Qwen2.5-7B-Instruct [24] to generate an answer $a \sim P_{\text{LLM}}(a|q, a^\star, c)$ where $q, a^\star$ is the geometric question and its groundtruth answer generated by a reasoning model (Gemini2 .5 Flash) corresponding to the caption $c^\star$ in advance. As encouraged by mainstream RL process, we check both the format and correctness of the answer, which is:

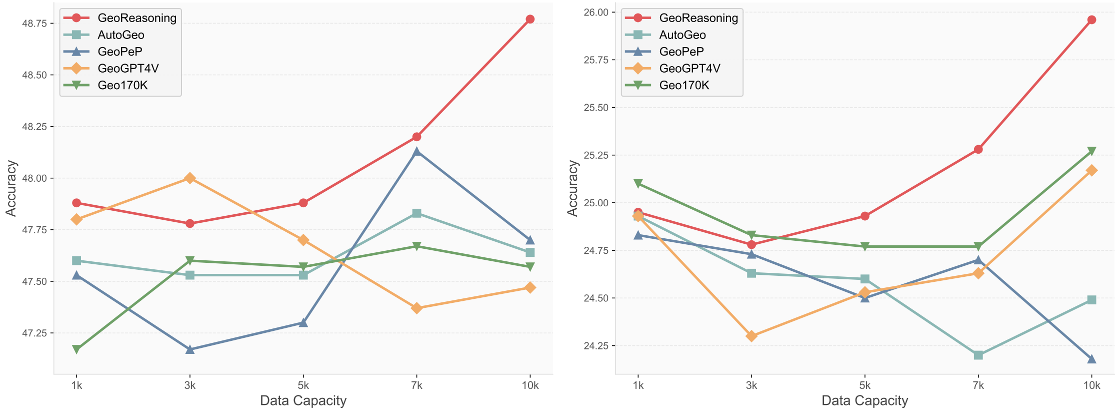
$$R_{\text{Reasoning}} = s_c \cdot \mathbb{I}(a = a^\star) + (1 - s_c) \cdot \mathbb{F}(a) \tag{4}$$

where $\mathbb{F}(\cdot)$ denotes the format checking function, and s_c indicate the weight of correctness, set as 0.9 in the experiments.

Caption reward To prevent reward sparsity during early training, we measure semantic relevance between c and the ground-truth caption $c^\star$ using ROUGE and BLEU-4, as shown in Eq. 5:

$$R_{\text{caption}} = w_r \cdot ROUGE(c, c^\star) + (1 - w_r) \cdot BLEU(c, c^\star) \tag{5}$$

where w_r represent the weight of ROUGE score, set as 0.7 in the experiments.



(a) MathVista (b) MathVerse

Figure 6: The accuracy of models SFTed on different capacities and mathematical datasets on downstream evaluation benchmarks: (a) MathVista. (b) MathVerse.

4 Experiments

We conduct a comprehensive evaluation of our proposed framework. First we show the experimental setup in detail in Section 4.1. Section 4.2 evaluates the in-domain performance of GeoReasoning-10K from the perspective of scaling property and reasoning capacity. Section 4.3 evaluates the out-of-domain performance of the RAFTed model and GeoReasoning-10K dataset, and shows case studies to explain the superiority of the proposed dataset. Section 4.4 illustrates the ablation study on training stages, and Appendix D serves as a complement about the ablation studies on various domains and hyperparameters.

4.1 Experimental Setup

Our experiments utilize Gemma3-4B [25], a commonly used lightweight multimodal architecture with strong mathematical reasoning capabilities as our base model. All the models are evaluated on MathVista [7] and MathVerse [6], two established mathematical reasoning benchmarks focusing on visual and mathematical problem-solving. The training and optimization pipeline contains two stages:

1. Cold start phase:

   In this stage we train each base model on the initial GeoReasoning-10K dataset for 1 epoch using standard supervised fine-tuning (SFT) with cross-entropy loss. The original learning rate of SFT is 1e-5, utilizing a cosine learning rate scheduler. We set 3% of the training steps as warm-up steps, where the learning rate grows linearly to the original learning rate 1e-5.

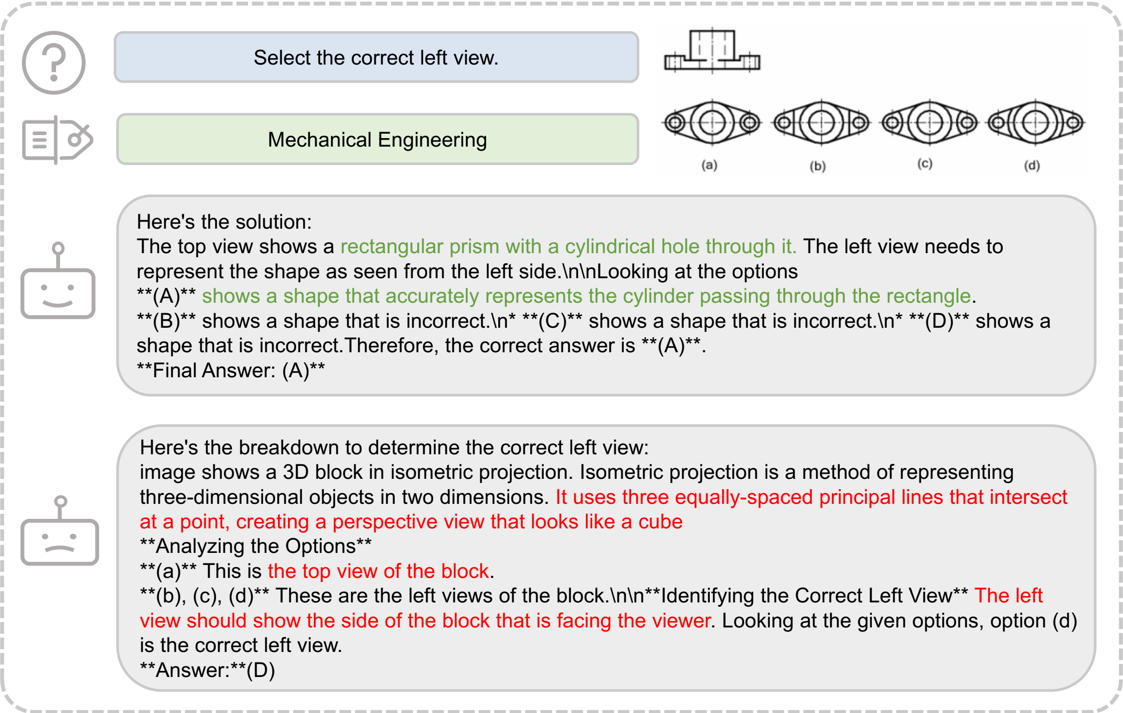
2. RAFT optimization phase:

   We conduct 5 refinement epochs, alternating between two sub-phases:

   - Caption Refinement: The current model generates 8 candidate captions for each image. Then we select the top 1 caption per image based on a composite reward considering both reasoning correctness and caption alignment. In the experiments, we set the weight of reasoning reward as 0.7 while that of caption reward as 0.3.
   - Model Retraining: Fine-tune the model on the updated dataset for 1 epoch using identical hyperparameters as the cold start phase.

4.2 In-Domain Performance of GeoReasoning-10K

In this section, we test the scaling property of our proposed dataset, compared to other commonly-used datasets(AutoGeo [16], GeoPeP [17], GeoGPT4 [26], and Geo170K [27]). Specifically, we randomly sample 1K, 3K, 5K, 7K, and 10K subsets from the original dataset. Then we train the baseline models on these datasets, after which evaluating on various benchmarks. The testing accuracy of models trained on various sizes of diverse datasets over multiple random seeds tested on MathVista and MathVerse are shown in Figure 6a and Figure 6b, respectively.


Select the correct left view.

Mechanical Engineering

Here's the solution:
The top view shows a rectangular prism with a cylindrical hole through it. The left view needs to represent the shape as seen from the left side.\n\nLooking at the options
(A) shows a shape that accurately represents the cylinder passing through the rectangle.
(B) shows a shape that is incorrect.\n* **(C)** shows a shape that is incorrect.\n* **(D)** shows a shape that is incorrect.Therefore, the correct answer is **(A)**.
Final Answer: (A)

Here's the breakdown to determine the correct left view:
image shows a 3D block in isometric projection. Isometric projection is a method of representing three-dimensional objects in two dimensions. It uses three equally-spaced principal lines that intersect at a point, creating a perspective view that looks like a cube
Analyzing the Options
(a) This is the top view of the block.
(b), (c), (d) These are the left views of the block.\n\n**Identifying the Correct Left View** The left view should show the side of the block that is facing the viewer. Looking at the given options, option (d) is the correct left view.
Answer:(D)

Figure 7: An engineering case, where the base model's answer is relatively general and the analysis of shape is not rigorous enough, while the model after training on GeoReasoning is more detailed and accurate in observing shape and has spatial reasoning ability.

Table 1: Accuracy of Gemma3-4B models trained on 10k random samples of AutoGeo, GeoPeP, GeoGPT4, Geo170K, and our dataset over multiple random seeds evaluated on various benchmarks.

	MathVerse	**MathVista**	**MMMU**
AutoGeo	24.5±0.4	47.6±0.5	43.5±0.5
GeoPeP	24.2±0.2	47.7±0.4	43.7±0.6
GeoGPT4	25.2±0.5	47.5±0.2	44.0±0.9
Geo170K	25.3±0.1	47.6±0.3	42.9±1.0
ours	**26.0**±0.3	**48.8**±0.2	**44.9**±0.7

Observed from Table 1, the model trained on GeoReasoning-10K has better reasoning ability compared to that trained on other caption datasets and even outperforms other multimodal reasoning datasets. This observation demonstrates a more pronounced ability of GeoReasoning-10K to enhance the model's reasoning capabilities.

As shown in Fig. 6, the model trained on GeoReasoning improves progressively with increasing dataset size on downstream benchmarks in general. Thus we can conclude that the model trained on GeoReasoning outperforms those trained on other mathematical datasets, indicating a more pronounced ability of GeoReasoning to enhance the model's reasoning capabilities. Besides, GeoReasoning has a much superior scaling law compared to other baselines.

4.3 Out-of-Domain Performance of GeoReasoning-10K

In this section, we evluate the generalization property of GeoReasoning-10K. Specifically, we evaluate the accuracy of the baseline model and the RAFTed model on a commonly-used benchmark MMMU [28]. The testing accuracy among 5 random seeds on all domains is shown in Table 2: Observed from Table 2, the trained model outperforms the baseline on most of the domains, indicating the generalization capacity of the model as well as the generalization property of the proposed dataset.

Besides, we compare the out-of-domain performance of models trained on GeoReasoning compared to that on various mathematical datasets, as shown in Table 1. The results reveal the stronger generalization property of GeoReasoning.

Table 2: Accuracy of models evaluated on all domains of MMMU, where 'A&D', 'Busi', 'Sci', 'H&M', 'Human', 'Tech' are short for 'Art and Design', 'Business', 'Science', 'Health and Medicine', 'Humanities and Social Science', 'Tech and Engineering', respectively.

	Overall	**A&D**	**Busi**	**Sci**	**H&M**	**Human**	**Tech**
Gemma3-4B	43.3±0.7	57.8±4.0	44.1±0.6	34.3±0.9	46.8±2.2	59.2±2.1	29.0±1.3
Gemma3-4B-RAFT	**44.9**±0.7	**60.2**±2.0	**44.5**±2.5	**36.0**±2.0	46.7±1.1	**60.0**±0.5	**32.9**±1.3

We also conduct qualitative analyses on representative examples from MathVista and MMMU to illustrate how our method enhances the model's ability across various domains. In particular, we select cases of different domains and compare the response of the RAFT optimized model and that of the base model. Figure 1 and Figure 7 present an example from the algebraic and engineering domains, respectively. Other examples in geometry, arithmetic, numeric, physics, and economics domains are shown in Appendix C.

In these cases, RAFT not only improves the reasoning ability in the geometry domain, but also generalize to other mathematical domains like arithmetic, algebraic, and numerical, even to non-mathematical domains like engineering, physics, and economics. This performance gain is hypothesized to stem from the reasoning abilities learned through the geometric captioning task.

4.4 Ablation Study

We validate the effectiveness of our Geo-Image-Textualization method by conducting ablation studies on training stages. Specifically, we implement the Cold Start and RAFT optimization pipelines on Gemma3-4B, generating refined models and datasets at each optimization stage. We then evaluate models of various stages on MathVista and MathVerse, where the results are shown in Table 3:

Table 3: Accuracy of Gemma3-4B models of various stages evaluated on MathVista and MathVerse.

	MathVerse	**MathVista**
Gemma3-4B	25.2	46.2
Gemma3-4B-Coldstart	25.9	47.6
Gemma3-4B-RAFT	26.1	49.4
Gemma3-4B-Coldstart-RAFT	**27.4**	**50.0**

It can be concluded that RAFT is effective for both the base model and the one after cold-start, and cold-start is essential for enabling the model to acquire task-related capabilities.

And the skills across diverse domains like geometry and arithmetic on various training stages are summarized in Table. 4 and Table. 5 in Appendix D.

5 Conclusion

In this paper, we propose **Geo-Image-Textualization**, a novel reinforcement learning-based framework designed to generate high-quality, geometry-centered multimodal data. Leveraging this framework, we construct **GeoReasoning-10K**, a new dataset aimed at bridging the gap between visual and linguistic modalities in the geometry domain. Extensive experiments on the MathVista and MathVerse benchmarks demonstrate that our framework significantly enhances the cross-modal reasoning capabilities of MLLMs when fine-tuned on **GeoReasoning-10K**, with improvements generalizing to other domains, even with non-geometric input images. These results highlight the potential of rule-based generation for cross-modal and cross-domain learning, pointing to a promising direction for future research at the intersection of multimodal reasoning.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923, 2025.

[2] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 24185–24198, 2024.

[3] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024.

[4] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. arXiv preprint arXiv:2412.16720, 2024.

[5] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. Advances in Neural Information Processing Systems, 37:8612–8642, 2024.

[6] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In European Conference on Computer Vision, pages 169–186. Springer, 2024.

[7] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. arXiv preprint arXiv:2310.02255, 2023.

[8] Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Ziyu Guo, Shicheng Li, Yichi Zhang, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, et al. Mavis: Mathematical visual instruction tuning with an automatic data engine. arXiv preprint arXiv:2407.08739, 2024.

[9] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. Advances in Neural Information Processing Systems, 37:95095–95169, 2024.

[10] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948, 2025.

[11] Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement. arXiv preprint arXiv:2503.17352, 2025.

[12] Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. arXiv preprint arXiv:2503.07536, 2025.

[13] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. arXiv preprint arXiv:2503.06749, 2025.

[14] Hanze Dong, Wei Xiong, Deepanshu Goyal, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and T. Zhang. Raft: Reward ranked finetuning for generative foundation model alignment. ArXiv, abs/2304.06767, 2023.

[15] Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. Solving olympiad geometry without human demonstrations. Nature, 625(7995):476–482, 2024.

[16] Zihan Huang, Tao Wu, Wang Lin, Shengyu Zhang, Jingyuan Chen, and Fei Wu. Autogeo: Automating geometric image dataset creation for enhanced geometry understanding. IEEE Transactions on Multimedia, 2025.

[17] Yanpeng Sun, Shan Zhang, Wei Tang, Aotian Chen, Piotr Koniusz, Kai Zou, Yuan Xue, and Anton van den Hengel. Mathglance: Multimodal large language models do not know where to look in mathematical diagrams. arXiv preprint arXiv:2503.20745, 2025.

[18] Junxiao Wang, Ting Zhang, Heng Yu, Jingdong Wang, and Hua Huang. Magicgeo: Training-free text-guided geometric diagram generation. arXiv preprint arXiv:2502.13855, 2025.

[19] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In Proceedings of the ieee/cvf conference on computer vision and pattern recognition, pages 13040–13051, 2024.

[20] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592, 2023.

[21] Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. xgen-mm (blip-3): A family of open large multimodal models. arXiv preprint arXiv:2408.08872, 2024.

[22] Renjie Pi, Jianshu Zhang, Jipeng Zhang, Rui Pan, Zhekai Chen, and Tong Zhang. Image textualization: An automatic framework for creating accurate and detailed image descriptions. arXiv preprint arXiv:2406.07502, 2024.

[23] Yiting Lu, Jiakang Yuan, Zhen Li, Shitian Zhao, Qi Qin, Xinyue Li, Le Zhuo, Licheng Wen, Dongyang Liu, Yuewen Cao, et al. Omnicaptioner: One captioner to rule them all. arXiv preprint arXiv:2504.07089, 2025.

[24] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. arXiv preprint arXiv:2412.15115, 2024.

[25] Clement Farabet and Tris Warkentin. Introducing gemma 3: The most capable model you can run on a single gpu or tpu. `https://blog.google/technology/developers/gemma-3/`, 2025.

[26] Shihao Cai, Keqin Bao, Hangyu Guo, Jizhi Zhang, Jun Song, and Bo Zheng. GeoGPT4V: Towards Geometric Multi-modal Large Language Models with Geometric Image Generation. arXiv e-prints, page arXiv:2406.11503, June 2024.

[27] Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, and Lingpeng Kong. G-LLaVA: Solving Geometric Problem with Multi-Modal Large Language Model. arXiv e-prints, page arXiv:2312.11370, December 2023.

[28] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. arXiv e-prints, page arXiv:2311.16502, November 2023.

A Question-Answer Pair Generation Prompt

The rule-and-LLM-based pipeline contains two stages. We first design a prompt that satisfies all the above conditions, using a relatively low temperature (0.2 in our experiments) to encourage the large model (Gemini 2.5 Flash) to generate initial questions based on the caption, while also labeling those that are inconsistent with the caption. For the inconsistent questions, we then switch to a different prompt, encouraging the model to incorporate additional information and formulate new questions accordingly, while increasing the temperature to 0.8. This process continues until a self-consistent question is generated for the first time.

The prompt of the first question generation stage is set as:

Prompt1

```
You are a helpful dataset processor. Please:
1. Generate a mathemetical question according to the given description of a
geometric image with the following requirements:
1.1 The problem should base on the given description, i.e., you must **NOT** generate
problems that are unrelated to the given description.
1.2 The problem difficulty should not be too low, such as determining some
information in the description.
1.3 It should also not be too hard, like introducing too much extra information, but
anyway you can introduce some extra information to form a good geometric problem.
1.4 You should **NOT** include or repeat any information in the description, and just
contain the real question. For example, if the description says: ‘Line segment AB
is present. The length of BA is 1.24.’, then when you generate the question, you
should not include the length of AB.
1.5 If the question is inconsistent with the given description, the final answer
should be ‘None’.
2. Answer the question you just provided, and express the final answer to two
decimal places. The final answer should be in \\boxed{{answer}}.

Description:
{description}
Generated Question:
{question}
Generated Response:
{response}
Final Answer:
\\boxed{{answer}}
```

The prompt of the question re-generation stage is set as:

Prompt1

```
You are a helpful dataset processor.  Please:  1.  Generate a mathemetical question
according to the given description of a geometric image with the following
requirements:
1.1 The problem should base on the given description, i.e., you must **NOT** generate
problems that are unrelated to the given description.
1.2 You can introduce some extra information to form a good geometric problem.
1.3 If you find that it is hard to generate some difficult questions, just give a
simple question.  For example, when the lengths of all four sides of a quadrilateral
are given, you can no longer assume it is a parallelogram or rectangle.  In such
cases, the problem may only allow for questions like asking for the perimeter, or
determining the length of a segment when a certain point divides a side into an
n-equal part, etc.
1.4 You should **NOT** include or repeat any information in the description, and just
contain the real question.  For example, if the description says:  ‘Line segment AB
is present.  The length of BA is 1.24.’, then when you generate the question, you
should not include the length of AB.
1.5 If the question is inconsistent with the given description, the final answer
should be ‘None’.
2.  Answer the question you just provided, and express the final answer to two
decimal places.  The final answer should be in \\boxed{{answer}}.
Description:
{description}
Generated Question:
{question}
Generated Response:
{response}
Final Answer:
\\boxed{{answer}}
```

B Experimental Setup and Dataset Information

B.1 Experimental Setup

To ensure consistency, we adopt the official evaluation codebases of both MathVerse and MathVista, using the GPT-4o-mini API to evaluate the performance of our MLLM. Specifically, following each benchmark's official setup, we use GPT-4o-mini to extract and assess the correctness of answers for MathVerse, and to extract answers for MathVista.

We evaluate MLLMs on MathVerse, MathVista and MMMU using A100 by VLLM. We employ Gemma3-4B as our base model and fine-tune it on Georeasoning-10K using 4 L20 GPUs. The training process is distributed using torchrun with the DeepSpeed ZeRO-3 optimization strategy.

B.2 Data Source

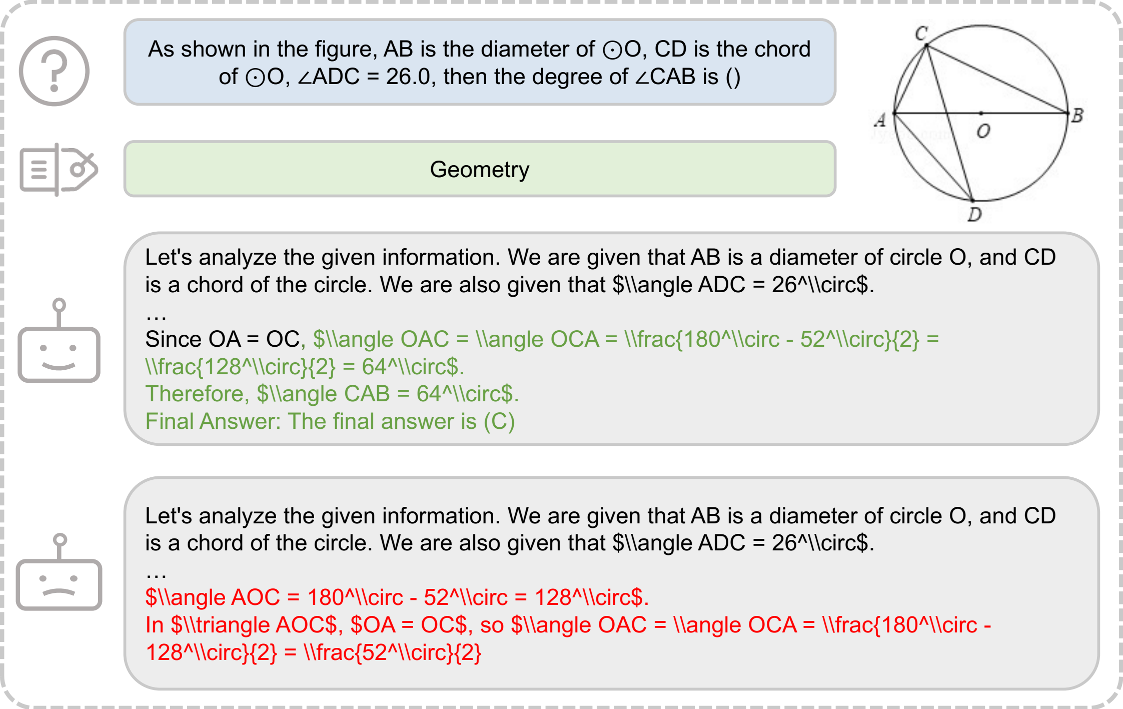
GeoReasoning-10K dataset is generated through rule-based methods and further refined using the RAFT framework. The question-answer pairs are generated by Gemini 2.5-Flash with a specific prompt.

B.3 License

- **GeoReasoning-10K** is released under the MIT License.
- **MathVerse** and **MathVista** are evaluated using their official codebases, which are publicly available under the Apache 2.0 License and MIT License, respectively.
- Our use of the **GPT-4o-mini API** for evaluation complies with OpenAI's API usage policies.
- All third-party datasets and models used in this work are under their respective licenses, and we ensure compliance with their terms of use.

C Case Studies

Additional mathematical cases that further demonstrate the generalization capability of our RAFT method are presented here. Figure 8, Figure 9, and Figure 10 correspond to the geometric, arithmetic, and numeric domains, respectively.



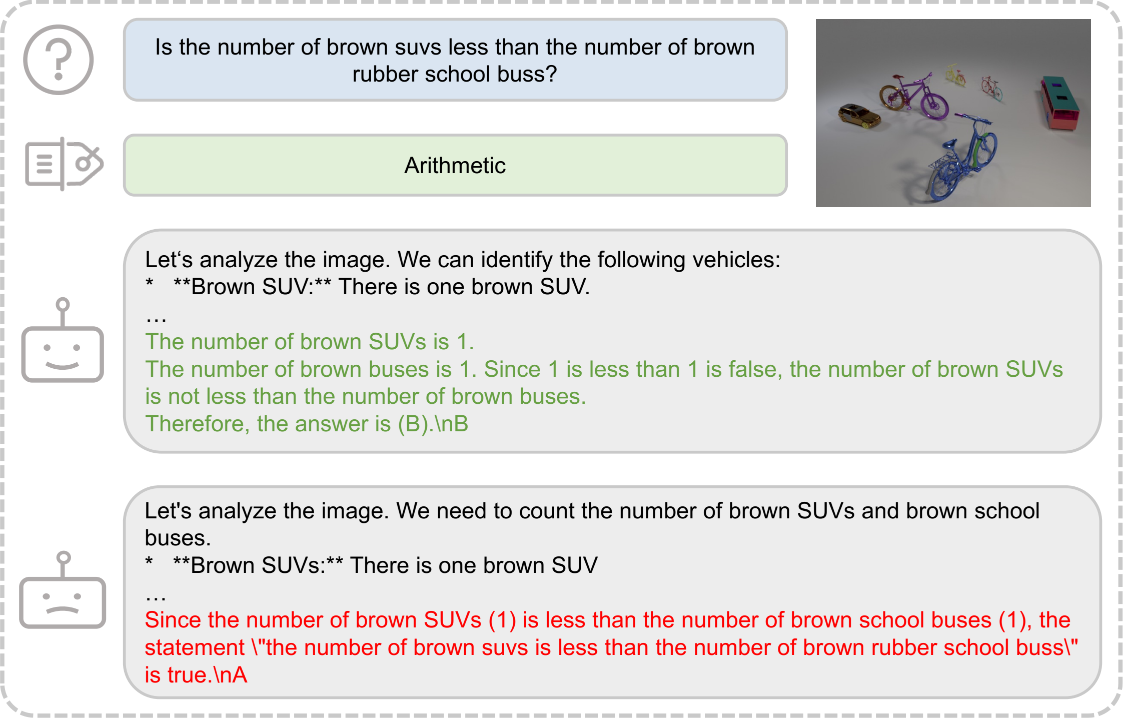
As shown in the figure, AB is the diameter of ⊙O, CD is the chord of ⊙O, ∠ADC = 26.0, then the degree of ∠CAB is ()

Geometry

Let's analyze the given information. We are given that AB is a diameter of circle O, and CD is a chord of the circle. We are also given that $\\angle ADC = 26^\\circ$.
…
Since OA = OC, $\\angle OAC = \\angle OCA = \\frac{180^\\circ - 52^\\circ}{2} = \\frac{128^\\circ}{2} = 64^\\circ$.
Therefore, $\\angle CAB = 64^\\circ$.
Final Answer: The final answer is (C)

Let's analyze the given information. We are given that AB is a diameter of circle O, and CD is a chord of the circle. We are also given that $\\angle ADC = 26^\\circ$.
…
$\\angle AOC = 180^\\circ - 52^\\circ = 128^\\circ$.
In $\\triangle AOC$, $OA = OC$, so $\\angle OAC = \\angle OCA = \\frac{180^\\circ - 128^\\circ}{2} = \\frac{52^\\circ}{2}

Figure 8: A geometric case.



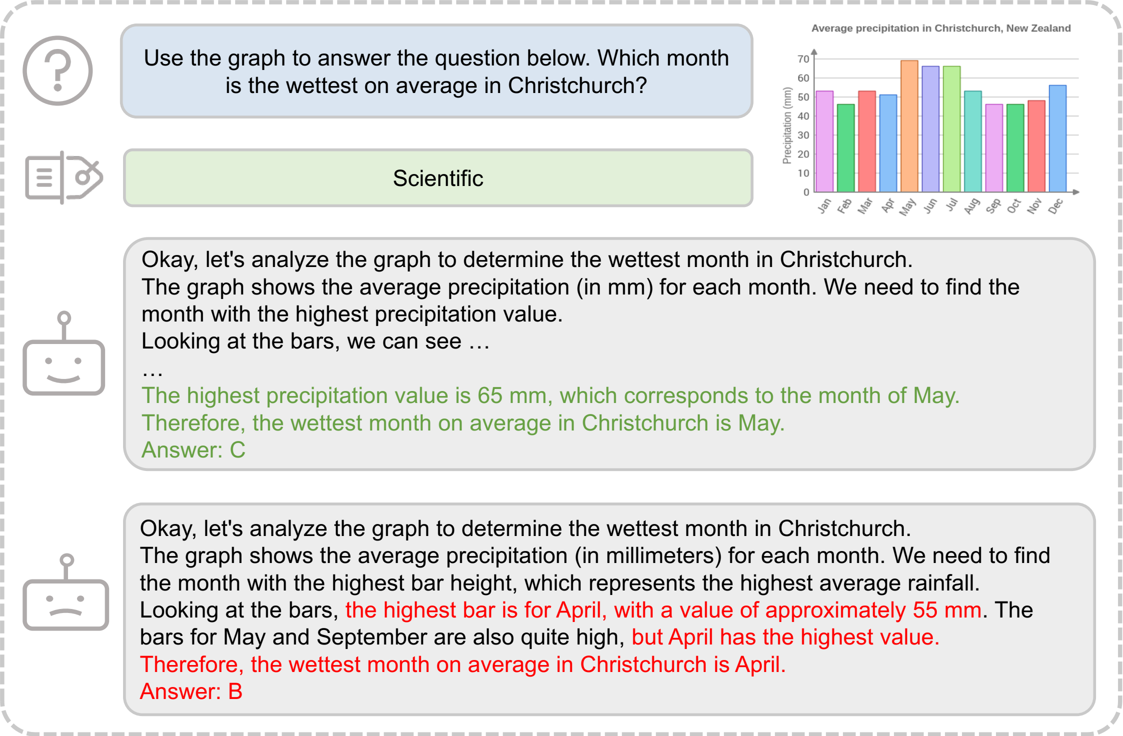
Is the number of brown suvs less than the number of brown rubber school buss?

Arithmetic

Let‘s analyze the image. We can identify the following vehicles:
* **Brown SUV:** There is one brown SUV.
…
The number of brown SUVs is 1.
The number of brown buses is 1. Since 1 is less than 1 is false, the number of brown SUVs is not less than the number of brown buses.
Therefore, the answer is (B).\nB

Let's analyze the image. We need to count the number of brown SUVs and brown school buses.
* **Brown SUVs:** There is one brown SUV
…
Since the number of brown SUVs (1) is less than the number of brown school buses (1), the statement \"the number of brown suvs is less than the number of brown rubber school buss\" is true.\nA

Figure 9: An arithmetic case.



Use the graph to answer the question below. Which month is the wettest on average in Christchurch?

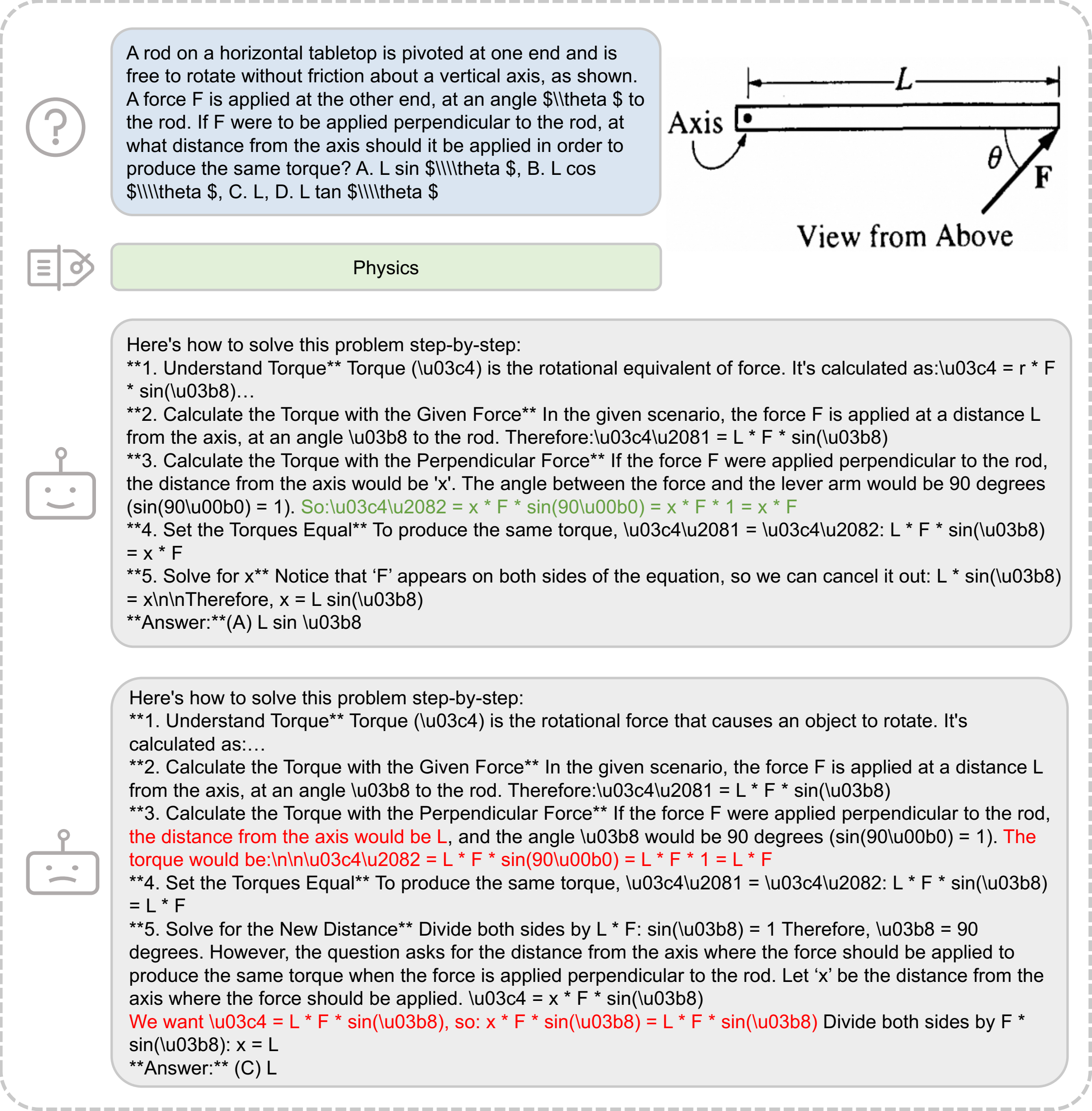
Scientific

Okay, let's analyze the graph to determine the wettest month in Christchurch.
The graph shows the average precipitation (in mm) for each month. We need to find the month with the highest precipitation value.
Looking at the bars, we can see …
…
The highest precipitation value is 65 mm, which corresponds to the month of May.
Therefore, the wettest month on average in Christchurch is May.
Answer: C

Okay, let's analyze the graph to determine the wettest month in Christchurch.
The graph shows the average precipitation (in millimeters) for each month. We need to find the month with the highest bar height, which represents the highest average rainfall.
Looking at the bars, the highest bar is for April, with a value of approximately 55 mm. The bars for May and September are also quite high, but April has the highest value.
Therefore, the wettest month on average in Christchurch is April.
Answer: B

Figure 10: A numeric case.

Figure 11 and Figure 12 exhibit non-mathematical cases of the physics and economics domains, respectively.



A rod on a horizontal tabletop is pivoted at one end and is free to rotate without friction about a vertical axis, as shown. A force F is applied at the other end, at an angle $\\theta $ to the rod. If F were to be applied perpendicular to the rod, at what distance from the axis should it be applied in order to produce the same torque? A. L sin $\\\\theta $, B. L cos $\\\\theta $, C. L, D. L tan $\\\\theta $

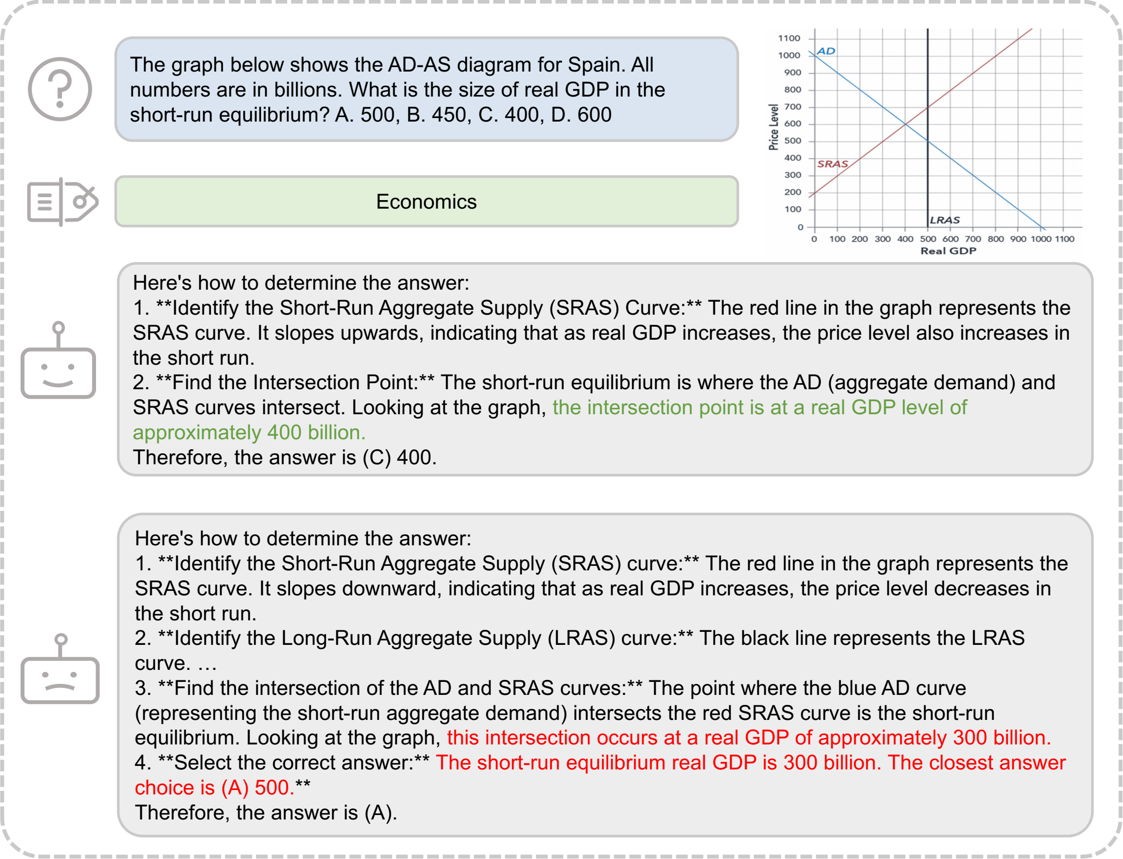
Physics

Here's how to solve this problem step-by-step:
1. Understand Torque Torque (\u03c4) is the rotational equivalent of force. It's calculated as:\u03c4 = r * F * sin(\u03b8)…
2. Calculate the Torque with the Given Force In the given scenario, the force F is applied at a distance L from the axis, at an angle \u03b8 to the rod. Therefore:\u03c4\u2081 = L * F * sin(\u03b8)
3. Calculate the Torque with the Perpendicular Force If the force F were applied perpendicular to the rod, the distance from the axis would be 'x'. The angle between the force and the lever arm would be 90 degrees (sin(90\u00b0) = 1). So:\u03c4\u2082 = x * F * sin(90\u00b0) = x * F * 1 = x * F
4. Set the Torques Equal To produce the same torque, \u03c4\u2081 = \u03c4\u2082: L * F * sin(\u03b8) = x * F
5. Solve for x Notice that 'F' appears on both sides of the equation, so we can cancel it out: L * sin(\u03b8) = x\n\nTherefore, x = L sin(\u03b8)
Answer:(A) L sin \u03b8

Here's how to solve this problem step-by-step:
1. Understand Torque Torque (\u03c4) is the rotational force that causes an object to rotate. It's calculated as:…
2. Calculate the Torque with the Given Force In the given scenario, the force F is applied at a distance L from the axis, at an angle \u03b8 to the rod. Therefore:\u03c4\u2081 = L * F * sin(\u03b8)
3. Calculate the Torque with the Perpendicular Force If the force F were applied perpendicular to the rod, the distance from the axis would be L, and the angle \u03b8 would be 90 degrees (sin(90\u00b0) = 1). The torque would be:\n\n\u03c4\u2082 = L * F * sin(90\u00b0) = L * F * 1 = L * F
4. Set the Torques Equal To produce the same torque, \u03c4\u2081 = \u03c4\u2082: L * F * sin(\u03b8) = L * F
5. Solve for the New Distance Divide both sides by L * F: sin(\u03b8) = 1 Therefore, \u03b8 = 90 degrees. However, the question asks for the distance from the axis where the force should be applied to produce the same torque when the force is applied perpendicular to the rod. Let 'x' be the distance from the axis where the force should be applied. \u03c4 = x * F * sin(\u03b8)
We want \u03c4 = L * F * sin(\u03b8), so: x * F * sin(\u03b8) = L * F * sin(\u03b8) Divide both sides by F * sin(\u03b8): x = L
Answer: (C) L

Figure 11: A physics case.

All these examples indicate that training on geometric caption tasks stimulates the reasoning capacity of models.



The graph below shows the AD-AS diagram for Spain. All numbers are in billions. What is the size of real GDP in the short-run equilibrium? A. 500, B. 450, C. 400, D. 600

Economics

Here's how to determine the answer:
1. **Identify the Short-Run Aggregate Supply (SRAS) Curve:** The red line in the graph represents the SRAS curve. It slopes upwards, indicating that as real GDP increases, the price level also increases in the short run.
2. **Find the Intersection Point:** The short-run equilibrium is where the AD (aggregate demand) and SRAS curves intersect. Looking at the graph, the intersection point is at a real GDP level of approximately 400 billion.
Therefore, the answer is (C) 400.

Here's how to determine the answer:
1. **Identify the Short-Run Aggregate Supply (SRAS) curve:** The red line in the graph represents the SRAS curve. It slopes downward, indicating that as real GDP increases, the price level decreases in the short run.
2. **Identify the Long-Run Aggregate Supply (LRAS) curve:** The black line represents the LRAS curve. …
3. **Find the intersection of the AD and SRAS curves:** The point where the blue AD curve (representing the short-run aggregate demand) intersects the red SRAS curve is the short-run equilibrium. Looking at the graph, this intersection occurs at a real GDP of approximately 300 billion.
4. **Select the correct answer:** The short-run equilibrium real GDP is 300 billion. The closest answer choice is (A) 500.**
Therefore, the answer is (A).

Figure 12: An economics case.

D Ablation Study (Supplement)

This section serves as a complement of Section 4.4, exhibit ing ablation studies on various domains, and hyperparameters of the reward function.

D.1 Ablation Study on Various Domains

We record the skills across diverse domains like geometry and arithmetic on various training stages in Table 4 and Table. 5:

Table 4: Accuracy of Gemma3-4B models at various stages tested on MathVista

	baseline	cold-start	raft-1	raft-2	raft-3	raft-4	raft-5
all	46.2	47.6	48.7	48.1	49.2	49.0	**50.0**
geometry	60.7	62.3	63.2	64.0	63.6	60.3	**64.0**
arithmetic	42.5	45.0	44.8	45.3	45.9	**47.6**	46.5
algebraic	59.1	60.5	62.3	62.3	62.3	59.1	**63.3**
numeric	26.4	31.9	29.9	31.3	31.3	31.9	**31.9**

It can be observed from Table 4 and Table 5 that the model after RAFT stages outperforms the base model across all domains. Specifically, the model achieves significant performance improvements across the arithmetic, algebraic, and numeric domains, with respective gains of 5.1%, 4.2%, and 5.5%. These results demonstrate the effectiveness of our approach in enhancing model performance as well as its generalization capability across different domains.

Table 5: Accuracy of Gemma3-4B models at various stages tested on MathVerse

	baseline	**cold-start**	**raft-1**	**raft-2**	**raft-3**	**raft-4**	**raft-5**
all	25.2	25.9	25.7	25.8	25.5	26.5	**27.4**
text dominant	32.0	35.5	35.1	35.2	35.1	36.5	**36.5**
text lite	25.9	27.4	28.2	**28.5**	27.4	26.6	26.3
vision intensive	24.0	24.8	24.4	24.4	23.1	26.1	**26.5**

D.2 Ablation Study on Hyperparameters

We evaluated the RAFTed models with various hyperparameters on MathVista and MathVerse, as shown in Table 6:

Table 6: Accuracy of RAFTed models with various hyperparameters evaluated on MathVista and MathVerse, where λ_r stands for the weight of reasoning reward.

	MathVista	**MathVerse**
λ_r = 1	49.8	**27.5**
λ_r=0.7	**50.0**	27.4
λ_r=0	48.9	27.5

As shown in Table 6, the reasoning reward plays a more important role in MathVista than MathVerse, indicating that the gain of genelization comes more from the helpness in solving the question other than comparison with captions.

In addition, it is observed in the result that the performance is not very sensitive to the selection of this hyperparameter, indicating the robustness of our RAFT method.

E Broader Impacts

The provided dataset pipeline and the generated dataset contribute to enhancing the generalizable reasoning abilities of multimodal large language models (MLLMs). In narrow domains, they are particularly effective for improving the geometric problem-solving capabilities of MLLMs, while in broader domains, they support the development of mathematical reasoning skills applicable to everyday scenarios. As the dataset is limited to geometric mathematical problems, it is considered safe for release and is unlikely to pose direct negative social impacts.